PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-069101
(43)Date of publication of application : 11.03.1997

(51)Int.Cl. G06F 17/27

(21)Application number : 07-223017
(22)Date of filing : 31.08.1995

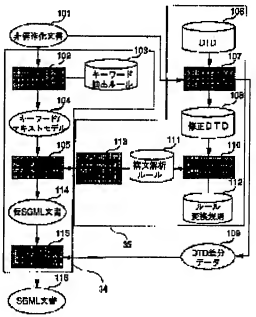
(71)Applicant : HITACHI LTD
(72)Inventor : SATO YOSHIFUMI
HINO MASATOSHI

(54) METHOD AND DEVICE FOR GENERATING STRUCTURED DOCUMENT

(57)Abstract:

PROBLEM TO BE SOLVED: To easily prepare a structured document matched with the logical structure of an individual document by executing conversion from a non- structured document to a structured document by the use of a rule directly prepared from previously set logical structure definition.

SOLUTION: A keyword extracting part 102 extracts a keyword expressing logical structure from a non-structured document 101 by the use of a keyword extraction rule 103 and generates a keyword/text model 104 expressing the document 101 by the keyword and two kinds of character string elements. A syntax analysis part 105 generated by a syntax analysis part automatic generation procedure 113 by referring to a syntax analysis rule 110 prepared by correcting/converting a DTD 106 executes syntax analysis for the model 104 and generates a temporary SGML document 114. An SGML document correcting part 115 corrects the document 114 by referring to DTD difference information 109 generated at the time of preparing the rule 110 and generates an SGML document 116 to be a final output.



LEGAL STATUS

[Date of request for examination]	04.08.2000
[Date of sending the examiner's decision of rejection]	07.09.2004
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	
[Date of registration]	
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－69101

(43)公開日　平成９年(1997)３月11日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/27			Ｇ０６Ｆ　15/20	５５０Ｆ ５５０Ａ	

審査請求　未請求　請求項の数９　ＯＬ　（全 15 頁）

(21)出願番号　特願平7－223017

(22)出願日　平成７年(1995)８月31日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　皆　俊史
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

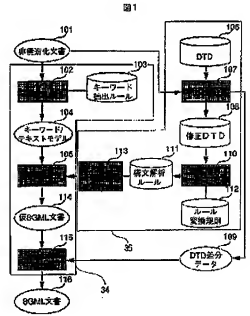
(72)発明者　鶴野昭利
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(74)代理人　弁理士　小川　勝男

(54)【発明の名称】　構造化文書生成方法および装置

(57)【要約】
【目的】非構造化文書から構造化文書への変換を、予め設定された論理構造定義から直接的に作成したルールを用いて行い、個々の文書の論理構造に即した構造化文書の作成を容易にする。
【構成】キーワード抽出部102は、キーワード抽出ルール103を用いて非構造化文書101から論理構造を表すキーワードを抽出し、非構造化文書101をキーワードとそれ以外の文字列の二種の要素で表現したキーワード／テキストモデル104を生成する。DTD106を修正・変換して作成した構文解析ルール110を参照して構文解析部自動生成手続き111が生成した構文解析部105は、キーワード／テキストモデル104に対する構文解析を行い、仮SGML文書114を生成する。SGML文書修正部115は、構文解析ルール作成時に生成されたDTD差分情報109を参照して仮SGML文書113を修正し、最終出力であるSGML文書116を生成する。

図1



1

【特許請求の範囲】

【請求項１】少なくとも入力表示装置，制御装置，および記憶装置を含み，文書の論理構造を規定する論理構造定義に基づき，構造が明示されていない文書である非構造化文書を前記入力表示装置により入力し，入力された非構造化文書を構造が明示されている文書である構造化文書に変換し，前記構造化文書を出力する構造化文書生成装置において，

予め与えられた第１の論理構造定義を，入力された非構造化文書の文書構造に適合させて修正して第２の論理構造定義を作成し，

作成された第２の論理構造定義を構成する記号および該記号の配列順序と１対１に対応すべく前記第２の論理構造定義を変形して前記第２の論理構造定義が規定する論理構造に適合した構文解析を行うための構文解析ルールを前記制御装置にて生成し，

生成された構文解析ルールに基づき，入力された前記非構造化文書から第１の構造化文書を生成し，

生成された第１の構造化文書を，前記第１の論理構造定義と前記第２の論理構造定義との差分データに基づき，前記第１の論理構造定義に従う形式に変換して第２の構造化文書を生成することを特徴とする構造化文書生成方法。

【請求項２】請求項１に記載の構造化文書生成方法において，

前記第１および第２の論理構造定義は，入力されるべき文書を構成する文字列間の相互関係を規定すべく配置された記号列であることを特徴とする構造化文書生成方法。

【請求項３】請求項１または２に記載の構造化文書生成方法において，

前記第１および第２の論理構造定義は，少なくとも，対応する文書中の各文字列の概念的な上下関係を表す論理構造を，前記各文字列の概念に当てた名称を所定の方法で記列して表現した記号列を含むことを特徴とする構造化文書生成方法。

【請求項４】請求項１または２に記載の構造化文書生成方法において，

前記非構造化文書から，文書中の文字列に係る所定のルールに基づきキーワードを抽出して，少なくともキーワードとして抽出された文字列とそれ以外の文字列とを含むキーワード／テキストモデルを生成し，

前記構文解析ルールを用いて前記キーワード／テキストモデルを前記第１の構造化文書に変換することを特徴とする構造化文書生成方法。

【請求項５】請求項４に記載の構造化文書生成方法において，

前記キーワードは，文字列の書式条件とキーワード名称とを対応づけたキーワード抽出ルールを参照して，前記非構造化文書中の文字列をいずれのキーワードであるか認識することによって抽出することを特徴とする構造化文書生成方法。

2

【請求項６】請求項５に記載の構造化文書生成方法において，

前記キーワード抽出ルールは，前記非構造化文書の出力書式定義が与えられている場合には，前記出力書式定義を所定のルールに基づき変換して作成することを特徴とする構造化文書作成方法。

【請求項７】請求項４に記載の構造化文書生成方法において，

同一の文字列領域から同一の文字列が異なる複数のキーワードとして抽出される場合には，前記制御装置が構文解析の成否を基準に前記複数のキーワードの中から適切なキーワードを選択することを特徴とする構造化文書生成方法。

【請求項８】請求項１または２に記載の構造化文書生成方法において，

前記構文解析ルールは，与えられたルール変換規則に基づき前記第２の論理構造定義を変換して生成された中間ルールに，構文解析時に解析された構文を明示するための手続きを埋め込んで生成することを特徴とする構造化文書生成方法。

【請求項９】少なくとも入力表示装置，制御装置，および記憶装置を含み，構造が明示されていない文書である非構造化文書を構造が明示されている文書である構造化文書に変換する構造化文書生成装置において，

前記非構造化文書のレイアウト情報と文字列情報とから，前記非構造化文書の論理構造の構成要素を表す文字列をキーワードとして抽出するキーワード抽出手段と，

与えられた第１の論理構造定義を修正して作成した第２の論理構造定義から，前記非構造化文書を前記第２の論理構造定義に適合する構造化文書に変換するルールを生成するルール生成手段と，

前記キーワード抽出部で抽出されたキーワードと前記ルール生成部で生成されたルールとを用いて前記構造化文書を生成する構造化文書生成部とを有することを特徴とする構造化文書生成装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は法規文書等一定の文書構造を有する文書の管理に関し，特に，文字認識やワードプロセッサ等の手段によって入力された，文書の構造を明示的に表す情報を含まない文書（以下「非構造化文書」と呼ぶ）を，文書の構造を明示的に表す情報を含む文書（以下「構造化文書」と呼ぶ）に変換する方法および装置に関する。

【０００２】

【従来の技術】構造化文書の形式の一つに，論理構造を明示的に表す情報をテキスト中に埋め込む方法がある。一般にユーザが作成した構造化文書（以下「文書インス

タンス」と呼ぶ)は、文書の論理構造を規定する論理構造定義を記述したファイルを指定する部分と、文書の内容を表すテキスト部とを含むことが多い。論理構造定義には、その文書の論理構造と、その構成要素を表すマーク(以下「タグ」と呼ぶ)とが定義される。このように、文書の論理構造化を行なう場合には、対象とする文書を活用するための論理構造定義が設定されている場合が多い。また、内容テキスト部には、論理構造定義内で定義されたタグを、そのタグに対応する論理構造の内容となる文字列が一意に定まるように挿入し、文書の論理構造を明示的に表現する。

【０００３】このようにして構造化された文書インスタンスを出力する際には、論理構造の各構成要素(以下「論理構造要素」と呼ぶ)をどのような書式で出力するかを規定する出力書式定義を記述したファイルを参照し、出力すべきイメージを生成する。この方法によると、文書インスタンスと出力書式定義とが独立しているために、出力に用いる個々の装置やシステムに関わらず文書インスタンスを交換することができる。

【０００４】また、こうした構造化文書における文字列の内容は、例えば「<著者名>」や「<タイトル>」というような、論理構造要素と一対一で対応するタグの挿入によって明示的に表現される。このため、構造化文書に対応した全文検索システム等のツールと組み合わせることにより、文書インスタンスの集合をそのままデータベースとして利用することができ、文書内容の追加・変更も容易となる。さらに、このデータベースの一部が障害等により喪失した場合、元の論理構造定義とデータベースである文書インスタンスとを照合することにより、データベース中に喪失部分があることを発見できる。

【０００５】こうした利点から、大量の文書を蓄積・利用する文書処理システムにおける文書管理形式として、構造化文書形式の採用が進んでいる。これに伴い、既存の紙面文書やワープロ入力文書などの非構造化文書を構造化文書へと変換する手法がいくつか提案されている。

【０００６】特開昭62-249270や「文書画像のODA論理構造化文書への変換方式(電子情報通信学会論文誌 D-II Vol.J76-DII No.11 pp.2274-2284)」においては、対象とする文書型の分野を限定し、その分野において共通性のある論理構造(以下「共通論理構造」と呼ぶ)及び論理構造認識ルールを用いて構造化文書の生成を行なう方法が提案されている。この方法では、まず「技術文書」「ビジネス文書」など対象とする文書の分野毎に、その分野で共通して用いることのできる論理構造を設定し、その論理構造に沿った論理構造認識を行なうためのルールを入手によって作成する。そして、そのルールを用いて非構造化文書を共通論理構造に則した文書インスタンスに変換する。さらに、共通論理構造において表現できない個々の文書の論理構造(以下「個別論理構造」と呼ぶ)固有の構成要素が存在する場合には、共通論理構造に沿った文書インスタンスを個別論理構造に沿った形に変換し直す。

【０００７】

【発明が解決しようとする課題】しかしながら、従来の方法では、論理構造認識を行なうための論理構造及びルールは対象になる文書の分野に依存するため、分野の異なる文書を扱う際には、その分野に対応したルールを新たに人手で生成する必要があり、その作業に多大な労力を要するという問題があった。

【０００８】また、従来の方法では、ある特定の分野における複数種類の文書に対して共通性の高いと思われる単一のルールを用いるため、そのルールは各々の文書に必ずしも適合するものではなく、個別論理構造固有の構成要素は直接認識することができなかった。そのために、論理構造認識後に、生成された文書インスタンスを個別論理構造に即した形式に変換し直す必要があった。具体的には、生成された文書インスタンスに含まれるタグの追加・変更・削除を行なうことになるが、一般にこの作業には多くの手間がかかり、結果として多大な労力を要するという問題があった。

【０００９】本発明は上記問題点に鑑み、ある特定の分野の複数の異なる種類の文書に従って適切な論理構造認識を行なう方法を提供することを目的とする。また、個別論理構造固有の構成要素を直接認識してその個別論理構造に即した形式の文書インスタンスを直接生成する方法を提供することを目的とする。

【００１０】

【課題を解決するための手段】上記目的を達成するため、本発明は、少なくとも入力表示装置、制御装置、および記憶装置を含み、文書の論理構造を規定する論理構造定義に基づき、構造が明示されていない文書である構造化文書を入力表示装置により入力し、入力された非構造化文書を構造が明示されている文書である構造化文書に変換し、構造化文書を出力する構造化文書生成装置において、予め与えられた第１の論理構造定義を、入力された非構造化文書の文書構造に適合させて修正して第２の論理構造定義を作成し、作成された第２の論理構造定義を構成する記号および該記号の配列順序と１対１に対応すべく第２の論理構造定義を変形して第２の論理構造定義の規定する論理構造に適合した構文解析を行うための構文解析ルールを前記制御装置にて生成し、生成された構文解析ルールに基づき、入力された非構造化文書から第１の構造化文書を生成し、生成された第１の構造化文書を、第１の論理構造定義と第２の論理構造定義との差分データに基づき、第１の論理構造定義に従う形式に変換して第２の構造化文書を生成することを特徴とする構造化文書生成方法とその構成とする。

【００１１】上記の構成において、非構造化文書から構造化文書への変換は、例えば、抽出したキーワードを手がかりにした構文解析によって論理構造認識を行なう場

5

文解析部を用いて行うことができる。構文解析部は、与えられた論理構造定義を構文解析ルール生成部によって構文解析ルールへと変換し、この構文解析ルールに構文解析部自動生成手続きを施すことによって生成される。

【００１２】ここで、構文解析部自動生成手続きとは、「aはB,C,…というパターンから構成される」というようなルールの集合を入力として、それらのルールに従った構文解析を実行するプログラムを出力するものであり、各ルールが成立した際に実行される特定の処理をルール内に記述することが可能である。このような構文解析部自動生成手続きとしては、例えば、UNIXに標準的に添付されるyaccが挙げられる。

【００１３】また、上記の構成において、特に同一の文字列領域の同一文字列が複数の異なるキーワードとして抽出される場合には、制御装置内の構文解析部が構文解析の成否を基準に当該複数種類のキーワードから適切な１つを選択する。

【００１４】構造化文書の具体的生成方法は以下の通りである。まず、キーワード抽出部が非構造化文書からキーワードを抽出し、対象とする文書をキーワードとそれ以外の文字列とを要素とする集合として抽象化したキーワード／テキストモデルを生成する。

【００１５】構文解析部は、キーワードテキスト／モデルに対して構文解析を行ない、構造化文書を生成するが、この構文解析部は以下の手順によって作成する。最初に、与えられた論理構造定義を非構造化文書が持つ論理構造に合わせて修正し、その差分情報を保持しておく。次に、構文解析ルール生成部が、修正後の論理構造定義を構文解析ルールへと変換する。このとき、各ルールが成立した際、すなわち各論理構造要素が検出された際に、検出された論理構造要素についての構造をキーワード／テキストモデルにおける該当部分に記録する処理プログラムを、構文解析ルールに組め込んでおく。そして、構文解析部自動生成手続きが、構文解析ルールに記述された構文解析処理を実現する構文解析部を生成する。

【００１６】以上の手続きによって生成された構文解析部は、キーワード抽出部の作成したキーワード／テキストモデルに対して構文解析を行ない、キーワード／テキストモデルに記録された構文解析結果を基に、修正後の論理構造定義に沿った仮構造化文書を生成する。構造化文書修正部は、論理構造定義の修正部に作成された論理構造の差分情報を参照して、修正前の論理構造定義に沿った構造化文書を出力する。

【００１７】

【作用】上記の構成によれば、論理構造とそれを認識するためのルールを、個々の文書に設定された論理構造定義から変換して作成するため、論理構造認識を行うための論理構造の設計およびルール作成に要する労力を軽減できる。また、個々の文書の論理構造定義を基に動的に作成した構文解析ルールを用いるため、共通論理構造を介することなく個別論理構造に即した構造化文書を直接生成することができ、構造化文書を共通論理構造に即した形から個別論理構造に即した形へと変換し直す必要がない。

【００１８】

【実施例】以下、図面を参照して本発明の一実施例を説明する。本実施例においては、構造化文書生成部が構文解析によって論理構造認識を行なうものとする。構造化文書形式としてSGML形式を採用し、文書の論理構造定義に相当する概念はSGMLの文書型定義であるDTD(Document Type Definition)であるとする。SGMLおよびDTDの処理内容や記述規則はISO(国際標準化機構)の標準規約であるISO 8879において規定されており、その詳細は文献「SGML入門」(MARTIN BRYAN著、アスキー出版局)に解説されている。また、本実施例においては、構文解析部自動生成手続きとして、UNIXに標準添付されているyaccを用いる。さらに、yaccが入力とする各ルールに対して、各々が成立した時点で実行される処理を付加する際の記述言語として、C言語を用いる。yaccの処理の詳細については文献「yaccとlexの使い方」(浅萩孝著、HBJ出版局)に、C言語については文献「プログラミング言語C」(B.W.カーニハン&D.M.リッチー著、共立出版)にそれぞれ解説されている。

【００１９】まず、本実施例のシステムの概要を説明する。図１９は本実施例に係る構造化文書生成システムのハードウエア構成図である。入力表示装置１は、操作者からの入力を受け付け、また、入力された非構造化文書や生成された構造化文書などを出力する装置であり、ディスプレイ、キーボード、マウス等から構成される。外部記憶装置２は、構造化文書生成に係る諸データを格納する装置であり、ハードディスク装置等により実現され、非構造化文書格納部２１、構造化文書生成ルール格納部２２、および構造化文書格納部２３を有する。制御装置３は、当該システムを構成する各装置の制御および構造化文書生成に係る情報処理等を行う装置であり、制御部３１、内部メモリ３２、および構造化文書生成部３３を有する。制御部３１は、非構造化文書格納部２１および構造化文書生成ルール格納部２２に格納されたデータを読み出して内部メモリ３２上に展開し、このデータを用いて内部メモリ３２上で構造化文書生成部３３の有する手続き、すなわち構文解析部生成手続き３４および構造化文書生成手続き３５を実行し、その結果生成される構造化文書を構造化文書格納部２３に格納する。構文解析部生成手続き３４は、構造化文書生成手続き３５の一部である構文解析部を生成する手続きである。構造化文書生成手続き３５は、構造化文書生成ルール格納部２２に格納された論理構造定義、キーワード抽出ルール、ルール変換規則等を用いて、非構造化文書格納部２１に格納された非構造化文書を構造化文書に変換する手続きである。構文解析部生成手続き３４および構

造化文書生成手続き303は公知のプログラミング言語等で記述できる。

【0020】次に、本実施例の処理概要を説明する。図1は、本実施例に係る構造化文書生成システムにおける構造化文書生成処理の流れを示すブロック図である。非構造化文書101は、ワードプロセッサや文字認識装置等によって作成される一次元的な文字列として電子化された文書情報であり、入力表示装置1によりシステムに入力される。キーワード抽出部102は、まず、キーワード抽出ルール103に従って非構造化文書からキーワードを抽出する。キーワードとは、非構造化文書101の論理構造を表現する文字列である。次に、非構造化文書101をキーワードとそれ以外の文字列とに分解し、これらを要素とする集合として抽象化したキーワード／テキストモデル104を生成する。構文解析部105は、構文解析ルール生成部110が作成した構文解析ルール111に記述された構文解析を実現し、論理構造認識を行なうものである。

【0021】構文解析部105の生成方法の概要は以下の通りである。まず、DTD修正部107において、DTD106を非構造化文書101の記述様式に沿うように修正してDTD108を作成し、その差分情報をDTD差分データ109として保持しておく。DTD106は、予め用意された標準的な論理構造定義であり、必ずしも入力された非構造化文書101に適合しているとは限らない。この修正は、非構造化文書101とDTD106とをシステム操作者が見比べた結果に基づいて行なう。構文解析ルール生成部110は、ルール変換規則112を参照して修正DTD108から構文解析ルール111を作成する。そして、本実施例における構文解析部自動生成手続きであるyacc113が、構文解析ルール111から、構文解析ルール111に記述された構文解析処理を実現する構文解析部105を生成する。

【0022】構文解析部105は、キーワード／テキストモデルに対する構文解析を行ない、論理構造を表すタグを付加して仮SGML文書114を生成する。これは、修正DTD108に沿った形で生成された文書インスタンスであるため、SGML文書修正部115が、DTD差分データ109を参照して仮SGML文書114を修正することにより、DTD106に沿ったSGML文書116を生成する。

【0023】次に、本実施例における各処理について詳細に説明する。図2は図1における非構造化文書101の例を示す。これは、法規を例に紙面文書に対して文字認識を行なった結果であり、論理構造を示す明示的な表記は存在しないが、文書の各構成要素はスペース等を用いて読み易いようにレイアウトされている。このようなテキスト形式の電子化文書を文書処理システムで活用するために、論理構造定義(DTD)が設定されている。図2の非構造化文書に対応するDTDの例を図3に示す。冒頭の301は、この論理構造定義が「条例」という名称であることを示す。302は、論理構造要素「条例」が、「公布」「例規番号」「題名」「本則」「附則」といった論理構造要素の並びによって構成されることを示す。「附則」にアスタリスク(“*”)が付いているのは、「附則」は任意個存在可能であることを意味する。303は、論理構造要素「公布」が「公布文」「公布年月日」「公布者」の並びによって構成されることを示し、307は「公布者」が「職名」と「氏名」から構成されることを示す。また、(#PCDATA)を構成要素とする304、305、307～310は、それぞれ「公布文」「公布年月日」「職名」「氏名」「例規番号」「題名」といった論理構造要素が、その内容を表す文字列を保持することを意味する。301から310までの論理構造をツリー状に表現したものを図4に示す。

【0024】本システムは、図2に示すような非構造化文書に対し、図3に示すようなDTDを直接的に利用した論理構造認識を行うことにより、そのDTDに従った構造化文書を生成する。

【0025】図1のキーワード抽出部102は、キーワード抽出ルール103を参照して非構造化文書101からキーワードを抽出し、キーワード／テキストモデル104を生成する。キーワード抽出ルール103の例を図5に示す。このルールは、キーワードとして抽出すべき論理構造要素名と、それを抽出するためのレイアウト情報及び文字列情報を記述した書式条件との組合せの集合である。図5においては、各行の先頭の項目がキーワードの名称であり、二番目以降の項目が書式条件である。図5における書式条件の記述要素の説明を図6に示す。これによれば、例えば図5における501は、キーワード「冒頭題名」の書式条件が、「行頭からスペース3文字の位置に文字「○」が存在し、それに任意長の文字列が続き、最後に文字列「条例」または文字列「規則」で行が終る。」という条件であることを意味する。また、502については、キーワード「公布年月日」の書式条件が、「行頭から任意個のスペースを置いて文字列「大正」または文字列「昭和」が存在し、その後は順に整数→「年」→整数→「月」→整数→「日」と続き、行が終る」という条件であることを意味する。

【0026】図1のキーワード抽出部102は、電子化文書の中にキーワード抽出ルールの書式条件に適合する文字列が存在するか否かを判定し、適合する場合にはその文字列をキーワードとして抽出する(キーワードの抽出例を図7に示す)。そして、対象文書をキーワードとそれ以外の文字列の集合として抽象化したキーワード／テキストモデル104を生成する。具体的には、キーワード間にキーワードに該当しない文字列が挟み込まれる場合、それをキーワード以外の文字列である「テキスト」とみなし、例えば図8に示すようなキーワード／テキストモデルを構成する。図8のキーワード／テキストモデルは、キーワード「冒頭題名」から始まり、その後キーワード「公布年月日」→キーワード「例規番号」→キーワード「公布文」→キーワード「題名」→キーワード「条番

号」と続く。キーワード「条番号」と次のキーワード「号番号」との間にキーワードでない文字列が挟まれるため、その部分が「テキスト」とみなされる。

【0027】ところで、文書中の同一の領域の同一文字列が複数箇所のキーワードとして抽出される場合がある。例えば、図7のキーワード抽出例において、一行目および二行目の文字列「○△△県水防信号規則」はそれぞれ「冒頭題名」および「題名」をキーワード名とするキーワードとして抽出されたものである。このような場合には、その領域からそれぞれのキーワードが抽出されたと仮定し、その仮定に対応したキーワード／テキストモデルを複数生成する。図8は、領域の競合するキーワード名「冒頭題名」および「題名」の中から「冒頭題名」を選択して生成したキーワード／テキストモデルの例である。この複数のキーワード／テキストモデルについては、後に説明する構文解析部105において構文解析が行われ、構文解析に失敗したものは不適切なキーワード／テキストモデルとみなされる。成功するものが複数存在する場合には、抽出されたキーワード数等の基準によって最適なものを選択し、最終的には最適なキーワード／テキストモデルに対応するSGML文書が一つだけ生成される。

【0028】図10の構文解析部105は、キーワード／テキストモデル104に対して、構文解析ルール111に従う構文解析処理を行なう。まず、構文解析ルール生成部107がDTD106を変換して構文解析ルール111を作成する過程を、図9を用いて説明する。

【0029】まず、DTD修正部107において、対象とする非構造化文書について設定されているDTD106の記述内容を、認識対象文書の記述様式に対応させるべく修正した修正DTD108を人手で作成し、その差分をDTD差分データ109として保持しておく。このような修正が必要になるのは、非構造化文書101の記述項目及びその記述順と、文書システムで利用する際に用いるDTD106における記述項目及びその記述順との間に矛盾が存在しうるためである。例えば、図3は、図2に示した非構造化文書101を活用するために用意されたDTD106である。しかし、図2の1行目の「○△△県水防信号規則」という冒頭の題名に対応する論理構造要素は、図3のDTD106の中には用意されていない。また、図3のDTD106においては、「公布文→公布年月日→例規番号→題名」という順に論理構造要素が並ぶことになっているのに対して、認識対象である図2の非構造化文書では、「公布年月日→例規番号→公布文→題名」という順番で各要素が並んでいる。

【0030】このような矛盾に対処するため、まず人手によって図10に示すような修正DTD108を作成する。網掛けで示した部分が修正を加えた部分である。この時、修正を加えた部分を明示的に表現するように、その部分が論理構造要素<CHANGE>で包含されるようにする。また、元のDTD106において修正された部分を、図11に示すようなDTD差分データ109として保持しておく。ここでも、修正された部分が論理構造<CHANGE>に包含されるようにする。

【0031】ただし、非構造化文書101に想定される論理構造とDTD106との間に矛盾が存在しない場合には、修正DTD108やDTD差分データ109を作成する必要はない。

【0032】必要に応じてDTD106に修正が加えられると、構文解析ルール生成部110は、図12に示すルール変換規則112に従ってルール変換906を実行し、修正DTD108に記述された論理構造に関する情報を中間vaccルール908へと変換する。中間vaccルールにおける各ルールは、「A：B C;」というようにコロン":"によって区切られた左辺と右辺から成り、右辺に記述された要素のパターンが存在する場合にルールが成立し、左辺の要素が構成される。例えば、「A：B C;」というルールの場合、「B C」というパターンが存在する時、要素Aが構成されることになる。

【0033】図10に示した修正DTD108を中間vaccルール908へ変換した例を図13に示す。例えば、図10における1091のルールを変換すると、図13の1301～1303に示したvaccルールに変換される。ここでは、1091の「目次?」と「前側*」が、図12における下から二つのルールによって、1301ではそれぞれ「opt0」「rep0」に置き換えられる。そして、「opt0」と「rep0」の定義がそれぞれ1302、1303に記述されている。

【0034】ところで、このような中間vaccルールを用いると、vaccの生成する構文解析部は構文解析の成否のみを出力し、キーワード／テキストモデルと論理構造要素との対応関係を出力しない。しかし、構文解析の結果を利用して構造化文書を生成するためには、各論理構造要素の認識に成功した際、すなわち中間vaccルールにおける各ルールが成立する時、対応するキーワード／テキストモデルに対して該当する論理構造要素に関する情報を付加する必要がある。そのために、構文解析ルール生成部110は中間vaccルール908に対して、キーワード／テキストモデルに情報付加処理を行なう言語のプログラムの埋め込み999を実行し、構文解析ルール111を生成する。構文解析ルール910の例を図14に示す。網掛けの部分が埋め込まれた言語の処理であり、ルールの右辺の構成要素に対応するキーワード／テキストモデルの情報を生成する処理を行なうものである。

【0035】図10において、vacc113は、生成された構文解析ルール111を入力として、構文解析ルール111に従った構文解析を行なう構文解析部105を生成する。DTD106から構文解析部105を生成する過程で人手を要するのは、論理構造定義を非構造化文書101の記述様式に合わせて変更し、差分DTDデータ109を生成する部分のみであり、残りの処理は自動的に行なわれる。

【００３６】図１０の構文解析部１０５は、キーワード／テキストモデル１０４に対して構文解析ルール１１１に従った構文解析を行なう。当該構文解析は、概して、次の２つのステップからなる。第一に、各テキストがどの論理構造要素に対応するのかを、隣接するキーワードに基づいて決定する。第二に、ツリー状表現のＤＴＤにおけるより下位の論理構造要素群をより上位の論理構造要素にまとめあげる。すなわち、当該ツリーの葉に相当する各論理構造要素が満たされたことから、それに対応する根に相当する論理構造要素が満たされたこととし、この根を新たな葉とみなして上記動作を繰り返し、最終的に最上位の論理構造要素を満たすことができたならば、入力されたキーワード／テキストモデルを当該ＤＴＤから生成された構文解析ルール１１１に適合するモデルとして認識する。

【００３７】本実施例においては、図８に示すキーワード／テキストモデルが入力され、例えば、「号番号」の“（１）”およびそれに続くテキスト“第１信号　警戒水位に達したことを知らせるもの”を読み取ると、まず、当該テキストが「号番号」なるキーワード“（１）”に隣接していることから、当該テキストは「号番号」に続く論理構造要素「号規定」に対応すると決定する。同様の処理を当該キーワード／テキストモデル中のすべてのテキストについて行なう。次に、例えば、上記の処理において“（１）”“第１信号　警戒水位に達したことを知らせるもの”という並びを読み取ったことにより、論理構造要素「号番号」および「号規定」が満たされたので、これらに対応する上位の論理構造要素「号」が満たされたと認識する。同様の処理により、隣接するすべての「号」、「号」の直前に位置する「条番号」、およびそれに続くテキスト“水防法～とする。”に対応する「初項規定」が満たされると、これらに対応する上位の論理構造要素「条」が満たされたと認識する。このようにして最終的に最上位の論理構造要素「条例」が満たされたならば、図８に示すキーワード／テキストモデル１０４を、図１０に示す修正ＤＴＤ１０８に基づき生成された構文解析ルールに適合するモデルとして認識する。

【００３８】構文解析の過程で、認識した論理構造要素に相当するキーワード及びテキストに対して、タグと一対一で対応する「タグ情報」を付加する。具体的には、構文解析によって、各論理構造要素に対応するキーワード／テキストモデルにおける要素が、「ｍ番目からｎ番目まで(m,nは整数でm<=n)」という形で得られるため、キーワード／テキストモデルのｍ番目の要素に、該当する論理構造がその要素から始まることを意味する「開始タグ情報」を付加し、ｎ番目の要素に対しては同様に「終了タグ情報」を付加する。この処理を、解析されたすべての論理構造要素について行う。そして、タグ情報の付加されたキーワード／テキストモデルから、キーワード及びテキストに相当する文字列の前後にSGMLのタグを付加した仮SGML文書１１４を出力する。仮SGML文書の例を図１５に示す。

【００３９】この例に示すように、タグ情報は「開始タグ情報」と「終了タグ情報」とから構成され、しかも「終了タグ情報」は「開始タグ情報」の近くにあるとは限らない。例えば、開始タグ情報「<条番号>」に対応する終了タグ情報「</条>」は図示されている範囲を超えてさらに下方に存在する。このため、仮SGML文書１１４が生成された段階で文書の構造を入手で修正しようとすると、対応する開始タグ情報と終了タグ情報とを文書全体にわたって探さなければならないため、多大な労力を要することになる。本発明においては、必要な修正をＤＴＤの段階で完了しているため、生成される仮SGML文書１１４は入力された非構造化文書１０１に即したものであり、上記のような修正は必要ない。

【００４０】同一の領域から複数のキーワードが抽出された場合には複数個のキーワード／テキストモデルが生成されるが、その場合には上記の構文解析処理をすべてのキーワード／テキストモデルに対して行う。誤ったキーワードを含むものは、構文解析に失敗する。構文解析に成功するキーワード／テキストモデルが複数個存在する場合には、例えば「抽出されたキーワード数が多い」ということを条件に、最適なキーワード／テキストモデルを選択し、それに対応した仮SGML文書を出力する。図７において非構造化文書の同一文字列から「冒頭題名」と「題名」の二つのキーワードが抽出された例を用いて説明すると、「題名」の方を選択して生成されるキーワード／テキストモデルは、構文解析に失敗する。これは、修正DTDの修正部分の一行目において、冒頭題名は条例の先頭に出現しうるが、「題名」は仮例の先頭には出現できないことが規定されているためである。そのため、「題名」に対応するキーワード／テキストモデルに対応する仮SGML文書は出力されない。一方、「冒頭題名」を選択して生成されるキーワード／テキストモデルは構文解析に成功するため、図１５に示すように、それに対応した仮SGML文書が出力される。

【００４１】DTD差分データ１０９が存在する場合には、そのデータを基にSGML文書修正部１１９が仮SGML文書１１４を修正する。具体的な処理内容を、図１６を用いて説明する。SGML文書修正部は、DTD差分データ１０９に記述された内容に対応する部分的なSGML文書である変更部インスタンス１６０２を作成する。このとき、論理構造の内容を表す文字列を意味する「#PCDATA」を対応する文字列に置換する必要があるが、その文字列を、仮SGML文書における変更部１６０３において同一名称の構成要素の内容を示す文字列によって置換する。例えば、変更部インスタンス１６０２において二つのタグ<公布文>及び</公布文>に挟まれた「#

13

PCDATA」は、仮SGML文書における変更部1603において同名のタグで挟まれた文字列である「△△県水防信号規則をここに公布する」に置換される。同様にして、二つのタグ<公布年月日></公布年月日>に挟まれた#PCDATAは文字列「昭和24年10月6日」に置換され、<例規番号></例規番号>に挟まれた#PCDATAは文字列「△△県規則第78号」に置換される。変更部インスタンス1602において タグ<組名></組名>に挟まれる#PCDATAのように、仮SGML文書における変更部1603に該当する同一名称の構成要素が存在しない場合には、強制的に「(なし)」という文字列を挿入することにする。

【００４２】以上の置換作業によって作成した変更部インスタンス1602を、図15の仮SGML文書114における変更部分、すなわち図15の例では二つのタグ<CHANGE>と</CHANGE>によって挟まれた部分と置換する。これにより、予め対象文書に対して設定されていたDTD106に即したSGML文書116として得る。SGML文書116の例を図17に示す。このSGML文書は個別論理構造を直接反映したものであるため、従来の方法のように文書インスタンスを個別論理構造へと変換する必要がない。

【００４３】ところで、認識対象文書に出力書式定義が与えられ、それに従った出力が行なわれる場合がある。また、論理構造定義が予め用意されている場合には、それを出力する際の出力書式定義が同時に与えられていることが多い。そして、特に定型文書の電子化については、従来の記述様式との互換性を保つため、出力書式定義が認識対象文書の記述様式に沿った形で作成されていることが多い。

【００４４】上記の実施例において、キーワード抽出ルール103は人手によって作成したものであるが、認識対象の文書に対して例えば図18に示すような出力書式定義が与えられており、かつ出力書式定義が認識対象文書の記述様式を基に作成されている場合や、出力書式定義に従って出力されている場合には、これを参照してキーワード抽出ルールを生成することも可能である。例えば図18は、論理構造要素「章番号」について「インデントは1cm」と規定されており、出力される文字列については「"第",CONTENT,"章"」と定められている。また、文字フォントには「huge-font」が用いられる。この時、例えばフォント「huge-font」の文字ピッチが1cmである場合には、行頭から2スペース1文字分の位置から文字列「第○章」が記述されることが分かる。このように、出力書式定義より、論理構造要素に対するキーワード抽出ルールを作成することが可能であり、例えば図4に示した形式で記述すると、「章番号 ^SPC1 "第" NUM1 "章" ;」というルールを得ることができる。

【００４５】認識対象文書の書式と出力書式定義に定義された書式が完全に一致しない場合には、キーワード抽出ルール103の人手による修正が必要になる。また、文字列情報が出力書式に記述されない場合には、人手によって文字列に関する書式条件を設定する必要がある。

14

【００４６】キーワード抽出ルール1036も構文解析ルール111と同様に、個々の文書に対して設定された情報を直接利用することによって、分野毎に共通した認識ルールを用いる従来方法では抽出できない個々の文書固有のキーワードが抽出可能となる。また、半自動生成を行うことにより、キーワード抽出ルール作成に要する労力を大幅に軽減することができる。

【００４７】

【発明の効果】本発明によれば、論理構造認識に用いる構文解析ルール111を対象文書に設定された論理構造定義から直接的に生成することにより、ルールの作成に要する労力を軽減することができる。また、個々の文書の論理構造定義に記述されている論理構造に従った構文解析によって文書インスタンスを生成するため、構文解析の結果得られる文書インスタンスを共通論理構造に沿った形から個別論理構造に沿った形へと変換し直す必要がない。

【図面の簡単な説明】

【図1】本発明の実施例に係る構造化文書生成システムの動作概要を説明するブロック図である。

【図2】非構造化文書の例を示した図である。

【図3】図2に示した文書に対して設定された SGML形式の論理構造定義であるDTD(一部)を示した図である。

【図4】図3に示したDTDの一部をツリー状に表現した図である。

【図5】キーワード抽出ルールの例(一部)を示した図である。

【図6】図5に示したキーワード抽出ルールにおける書式条件の記述要素を説明した図である。

【図7】キーワードの抽出例を示した図である。

【図8】キーワード/テキストモデルの例を示した図である。

【図9】構文解析ルール生成部の動作概要を説明するブロック図である。

【図10】修正DTDの例(一部)を示した図である。

【図11】DTD部分データの例を示した図である。

【図12】構文解析ルール生成部がDTDをyaccルールに変換する際に参照する変換規則を示した図である。

【図13】中間yaccルールの例(一部)を示した図である。

【図14】構文解析ルールの例(一部)を示した図である。

【図15】仮SGML文書の例(一部)を示したである。

【図16】SGML文書修正部の処理例を示した図である。

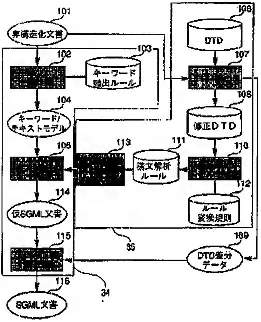
【図17】実施例に述べる方法によって最終的に得られるSGML文書の例(一部)を示した図である。

【図18】出力書式定義の例(一部)を示した図である。

【図19】本実施例に係る構造化文書生成システムのハードウエア構成を示した図である。

【図1】

図1



【図6】

図6

要素	意味
^	行頭.
$	行末.
" ..."	文字列. ダブルクォーテーションで挟んで記述.
NUM{数字}	NUM1→数字、NUM2→カタカナ、NUM3→英小文字.
SPC{数字}	{数字}個分のスペース.
!	行頭からの任意個のスペース.
?	任意の一文字.
÷	一文字以上の任意の文字列.
({要素}|{要素})	要素のOR.

【図7】

図7

管理題名　[　　　○△△県水防信号規則]
題名　[　　　○△△県水防信号規則]
公布年月日　[昭和24年10月6日]
例規番号　[△△県規則第78号]
公布文　[　△△県水防信号規則をここに公布する。]
題名　[　　　△△県水防信号規則]
条番号　[第1条　]
号番号　[　（1）　]
号番号　[　（2）　]
号番号　[　（3）　]
号番号　[　（4）　]
条番号　[第2条　]
附則　[　　附　則]
別表番号　[別表]

【図2】

図2

○△△県水防信号規則

昭和24年10月6日
△△県規則第78号

△△県水防信号規則をここに公布する。
　△△県水防信号規則
第1条　水防法（昭和24年6月法律第193号）第13条第1項の規定による水防信号は、次に掲げるものとする。
（1）　第1信号　警戒水位に達したことを知らせるもの
（2）　第2信号　水防団員及び消防機関に属する者の全員が出動すべきことを知らせるもの
（3）　第3信号　当該水防管理団体の区域内に居住する者が出動すべきことを知らせるもの
（4）　第4信号　必要と認める区域内の居住者に避難のため立ち退くべきことを知らせるもの
第2条　水防信号は、別表に定める区分及び方法に従って発する。
　附　則
この規則は、公布の日から施行し、昭和24年9月3日から適用する。
別表
水　防　信　号

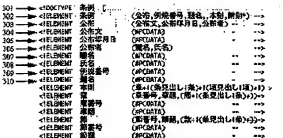
【図11】

図11

```
<!DOCTYPE CHANGES [
<!ELEMENT CHANGES   (公布,例規番号,題名)      -->
<!ELEMENT 公布      (公布文,公布年月日,公布者)  -->
<!ELEMENT 公布文    (#PCDATA)                  -->
<!ELEMENT 公布年月日 (#PCDATA)                  -->
<!ELEMENT 公布者    (職名,氏名)                -->
<!ELEMENT 職名      (#PCDATA)                  -->
<!ELEMENT 氏名      (#PCDATA)                  -->
<!ELEMENT 例規番号  (#PCDATA)                  -->
<!ELEMENT 題名      (#PCDATA)                  -->
]>
```

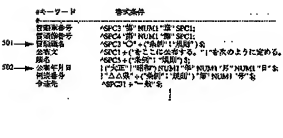
【図３】

図3



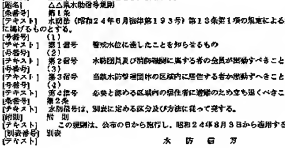
【図５】

図5



【図８】

図8



【図１３】

図13

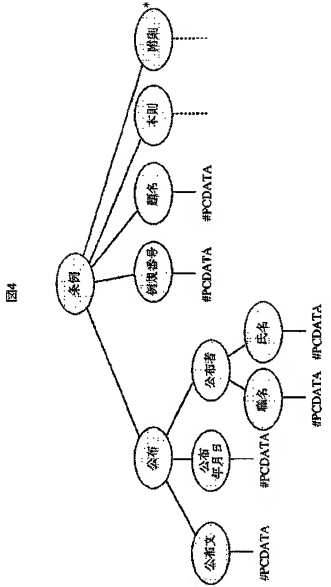
【図１８】

図18

```
(条番号(
    (インデント:　1cm)
    (フォント　:　Hoge_font)
    (出力文字列:　"第" CONTENT "条")
))
(節番号(
    (インデント:　1.5cm)
    (フォント　:　Large_font)
    (出力文字列:　"第" CONTENT "節")
))
```

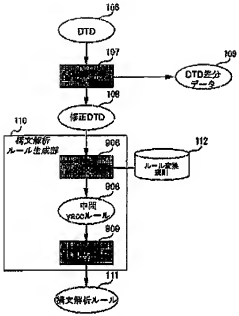
【図4】

図4


本例
公布
公布文
#PCDATA
公布年月日
#PCDATA
公布者
氏名
#PCDATA
職名
#PCDATA
制度番号
#PCDATA
題名
#PCDATA
本則
附則
*

【図９】

図9



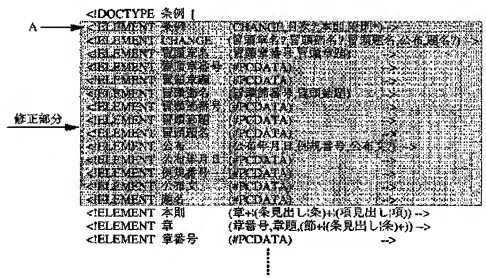
【図１２】

図12

DTDにおける記述形式	意味		yaccルール
(A , B)	全ての要素は指定された順序に現われる。	➡	(A B)
(A & B)	全ての要素はどんな順序に現われてもよい。	➡	(A B) | (B A)
(A | B)	一つの要素だけが現われる。	➡	(A | B)
A+	1回以上出現し、繰り返し指摘可能な要素。	➡	Aplus ただし Aplus: A | (Aplus A);
A?	1回まで、または出現しなくてもよい要素。	➡	Aopt ただし Aopt: | A;
A*	2回以上出現可能でかつ出現しなくてもよい要素。	➡	Aasp ただし Aasp: | (Aasp A);

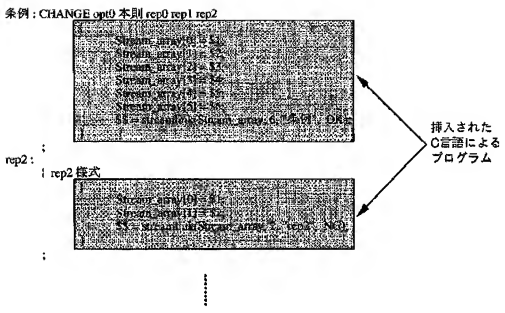
【図１０】

図10



【図14】

図14


条例：CHANGE opt0 本則 rep0 rep1 rep2
挿入された
C言語による
プログラム
rep2：
rep2 様式

【図16】

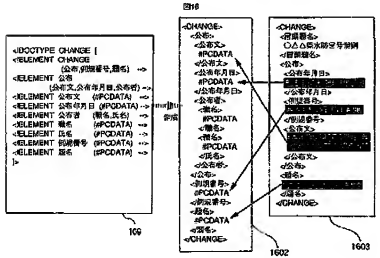
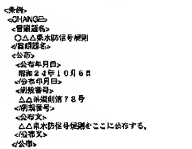


図16
<!DOCTYPE CHANGE [
<!ELEMENT CHANGE
(公布,(附則番号,題名) ?)>
<!ELEMENT 公布
(公布文,公布年月日,公布者) +>
<!ELEMENT 公布文 (#PCDATA) >
<!ELEMENT 公布年月日 (#PCDATA) >
<!ELEMENT 公布者
(題名,氏名) *>
<!ELEMENT 題名 (#PCDATA) >
<!ELEMENT 氏名 (#PCDATA) >
<!ELEMENT 附則番号 (#PCDATA) >
<!ELEMENT 題名 (#PCDATA) >
]>
<CHANGE>
<公布>
<公布文>
#PCDATA
</公布文>
<公布年月日>
#PCDATA
</公布年月日>
<公布者>
<題名>
#PCDATA
</題名>
<題名>
#PCDATA
</題名>
</公布者>
</公布>
<附則番号>
#PCDATA
</附則番号>
<題名>
#PCDATA
</題名>
</CHANGE>
<CHANGE>
<附則番号>
○△□※☆§¶†‡
</附則番号>
<公布>
<公布年月日>
</公布年月日>
</公布者>
</公布>
<題名>
</CHANGE>
109
1602
1603

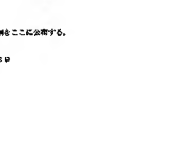
【図１５】

図15



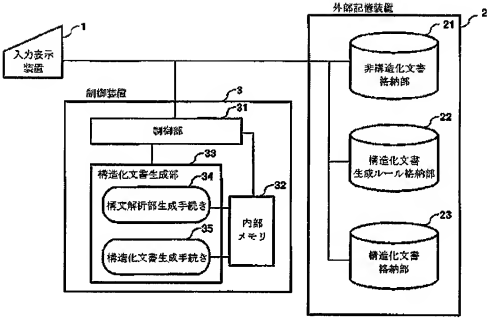
【図１７】

図17



【図19】

図 19


1
入力表示装置
制御装置
3
31
制御部
33
構造化文書生成部
34
構文解析部生成手続き
35
構造化文書生成手続き
32
内部メモリ
外部記憶装置
2
21
非構造化文書格納部
22
構造化文書生成ルール格納部
23
構造化文書格納部